論文の内容の要旨

論文題目　　音声認識のための高精度，コンパクトな音響モデルの研究
氏　　名　　磯 健一

近年，コンピュータの演算・記憶能力の増大と，大規模に収集された音声データベースの整備を追い風として，隠れマルコフモデル（Hidden Markov Model，HMM）を用いた統計的な音声認識技術の研究開発が精力的に進められている．HMM を用いると，大規模音声データベースから不特定話者の音声に関する統計的なモデルを自動学習することができる．これにより以前には困難であった不特定話者の大語彙音声認識が実験室環境で動作するようになってきた．しかし音声認識の本格的な実用化に向けてはまだ多くの課題が残されている．すなわち実用場面で十分な認識精度を実現するためには各部の高精度化や，話者・環境変化に対する頑健性，少ない CPU・メモリリソースで動作するためのコンパクト化，などが重要な研究課題である．大語彙連続音声認識の認識精度は，ていねいで協力的な発声に対して 9 割を超えるレベルに到達しつつあるが，多くのアプリケーションにおいて必ずしもまだ十分な精度ではない．話者による性能のバラツキも大きく，マイクロホン不一致，周囲雑音・残響などにも敏感である．また高速 CPU と大容量メモリを搭載した PC では動作可能であるが，携帯端末や家電，自動車などに組み込むことは容易ではない．このような背景の下で，本研究ではとくに音響モデルに焦点をあてて，これら 3 つの課題（高精度化，話者適応化，コンパクト化）について検討を行った．

音声認識精度を改善するための第 1 の課題として，音声時系列パターンのモデルである音響モデル（HMM）の高精度化を試みた．HMM は音声を定常的な音声区間（状態とよぶ）の連鎖としてモデル化しており，同一の状態内では，音声の特徴ベクトルはお互いに独立に同一の確率分布から出力されたものと仮定されている．すなわち時間的に隣り合う音声特徴ベクトルのあいだの相関がモデルには考慮されていない．現実の音声パターンは，調音結合に代表されるように隣接する特徴ベクトルのあいだには強い相関が存在している．より高い精度の音声認識を実現するためには，このような音声パターンの動的な性質（隣接する特徴ベクトルの相関など）を適切に音響モデルに取り入れることが重要である．
そこで多層パーセプトロンによる音声特徴ベクトル時系列の非線形予測を用いたニューラル予測モデル（Neural Prediction Model，NPM）を提案し，その認識・学習アルゴリズム

を定式化した．不特定話者離散数字認識実験によって，従来方式である DP マッチングや線形予測に比べて優れていることを示した．また多層パーセプトロンをパターン識別に用いた DNN（Dynamic Programming Neural Network）に比べても良好な認識性能が得られることを確認し，多層パーセプトロンをパターン予測に用いることの有効性を示した．さらに特定話者大語彙単語認識に適用するために，単語より小さな認識単位である半音節単位の NPM を定式化し，その調音結合に対する追従性を高めるために後ろ向き予測機能を追加し，その効果を確認した．また NPM では状態ごとに別々の予測器を用いてモデルを構成しているが，それらを 1 つに共有化して，駆動力ベクトルによって切り替える方式も検討し，英語アルファベット認識によってその効果を検証した．さらにその学習アルゴリズムに識別学習（事後確率最大化学習）を導入して性能改善を確認した．NPM は混合ガウス分布 HMM として解釈すると，HMM の各ガウス分布平均ベクトルが定数値ではなく，予測器により入力音声に対応して時々刻々と変化するモデルに相当し，HMM に局所的時間相関の表現能力を加えるという初期の目標に対して一つの解決策を与えることができた．

第 2 の課題として不特定話者 HMM の新しい話者への適応化（話者適応化）に関する研究を行った．不特定話者 HMM の学習に用いられる大規模音声データベースは必ずしも想定されるすべてのパターン変動を含んでいるわけではない．話者の変動や周囲雑音，マイクロホンや回線などの伝送特性の変動などは，きわめて多岐にわたるため，そのすべてを音声データベースに含めることは現実的には不可能である．そこで代表的なパターンをできるだけ多く集めて音声データベースを構築し，それらを用いて不特定話者 HMM を学習している．そして音声認識を使用する場面において，新たに少量の音声データを集めて，不特定話者 HMM をその場面に適応化させる適応化技術の重要性が増している．現在広く用いられている話者適応化技術（MLLR 法や MAP 法）では，HMM のモデルパラメータに対して不特定話者 HMM から得られる知見に基づいて制約を導入して，新しい話者の少量発声からでも安定にパラメータが推定できるようにしている．しかしこれらの手法では，大規模な音声データベースに含まれている話者変動の平均的な知見のみを利用しており，多数の個別話者の事例は有効に活用されていない．より少ない発声を用いて安定・高精度に新話者に適応化する話者適応化技術を実現するためには，大規模音声データベースを平均的な不特定話者 HMM の学習に用いるだけでなく，そのデータベースに含まれている多数の話者変動の事例を統計的にモデル化して適応化方式に取り入れることが重要である．
そこで大規模音声データベース中の話者変動の事例をモデル化した話者適応化方式である「EigenVoice 法」と，話者適応化用音声データの分量に応じて適応化すべきモデルパラメータ数を自律的に制御する「自律的モデル複雑度制御法（AMCC 法）」を融合した新しい話者適応化方式として階層的 EigenVoice 法（HEV 法）を提案した．不特定話者の混合ガウス分布 HMM のガウス分布を木構造にクラスタリングし，大規模な音声データベースから学習した多数の特定話者 HMM から得られる話者変動事例を用いて，木構造の各ノードご

とに話者変動の固有ベクトルを自動推定する学習アルゴリズムを定式化した．また話者適応化に用いることができる発声量に応じて木構造のノードを自律的に選択しながら話者適応を行う適応アルゴリズムとして最尤推定に基づく MLED 法と推定パラメータの事前分布を用いた MAPED 法を定式化した．HMM を用いた日本語大語彙連続音声認識（語彙 8 万語ディクテーション）に適用して，適応化用の 5 文発声で，従来法（MLLR 法や AMCC 法）の 50 文発声に相当する認識性能が得られることを確認し，話者変動事例知識活用の有効性を示すことができた．

第 3 の課題として，大語彙音声認識の実用化に向けたコンパクト化の検討を行った．不特定話者の大語彙連続音声認識を可能にする HMM は，大量のモデルパラメータの格納と，それらを用いた大規模な演算処理を必要とする．このような HMM を携帯情報端末や携帯電話などに組み込んで動作させるためには，HMM モデルパラメータのコンパクトな表現方法や，それらを用いた高速なパターンマッチング方法の開発が必須と考えられる．そこで HMM による不特定話者大語彙連続音声認識（ディクテーション）を実用化に向けてコンパクト化する方式として，音響モデルのコンパクト化方式（MDL 基準による混合ガウス分布数削減，ガウス分布対角共分散行列の共有化，ガウス分布木構造化による高速確率計算）の有効性を評価した．それらを，言語モデルのコンパクト化方式（クラス化），サーチのコンパクト化方式（音素木構造辞書の動的なトライフォン展開，ガベージコレクション，言語スコア計算再利用法）などと統合して日本語大語彙ディクテーション（語彙 5000 語）を市販 PDA 上に実装し，メモリ使用量約 4Mbyte で実時間応答（単語誤り率 8.4%）することを確認した．

以上のように本研究により，HMM による音声認識方式の新たな研究方向として，非線形予測の導入による高精度化，話者変動事例のモデル化による頑健性強化，の可能性が開かれた．また現在の HMM による不特定話者大語彙連続音声認識技術の一つの到達点として，それらをメモリ使用量や演算量に制約がある PDA 上で実用的な性能で動作させる実装アルゴリズムを示すことができた．